medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

VGsim: scalable viral genealogy simulator for global pandemic

Vladimir Shchur*†, Vadim Spirin†, Dmitry Sirotkin†, Evgeni Burovski†, Nicola De Maio‡, and Russell Corbett-Detig†,¶

†HSE University, Russian Federation
‡European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
¶Department of Biomolecular Engineering and Genomics Institute, UC Santa Cruz, California 95064

November 2021

Abstract

Accurate simulation of complex biological processes is an essential component of developing and validating new technologies and inference approaches. As an effort to help contain the COVID-19 pandemic, large numbers of SARS-CoV-2 genomes have been sequenced from most regions in the world. More than 5.5 million viral sequences are publicly available as of November 2021. Many studies estimate viral genealogies from these sequences, as these can provide valuable information about the spread of the pandemic across time and space. Additionally such data are a rich source of information about molecular evolutionary processes including natural selection, for example allowing the identification of new variants with transmissibility and immunity evasion advantages. To our knowledge, there is no framework that is both efficient and flexible enough to simulate the pandemic to approximate world-scale scenarios and generate viral genealogies of millions of samples. Here, we introduce a new fast simulator `VGsim` which addresses the problem of simulation genealogies under epidemiological models. The simulation process is split into two phases. During the forward run the algorithm generates a chain of population-level events reflecting the dynamics of the pandemic using an hierarchical version of the Gillespie algorithm. During the backward run a coalescent-like approach generates a tree genealogy of samples conditioning on the population-level events chain generated during the forward run. Our software can model complex population

*Corresponding author: vshchur@hse.ru

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

structure, epistasis and immunity escape. The code is freely available at `https://github.com/Genomics-HSE/VGsim`.

1 Introduction

The unprecedented world-wide effort to produce and share viral genomic data for the ongoing SARS-CoV-2 pandemic allows us to trace the spread and the evolution of the virus in real time, and has made apparent the need for improved computational methods to study viral evolution [1]. These data yield important insights into the effects of population structure [2–5], public health measures [6, 7], immunity escape [8,9], and complex fitness effects [10,11]. It is essential that we also have tools to accurately simulate viral evolutionary processes so that the research community can validate inference methods and develop novel insights into the effects of such complexities. However, there are no software packages capable of simulating the scale and apparent complexity of viral evolutionary dynamics during the SARS-CoV-2 pandemic.

Pandemic-scale datasets impose technical problems associated with the scalability and memory usage of computational methods. There is already substantial progress in building scalable simulators and data analysis methods for human genome data. The current state-of-the-art human genome simulator msprime [12] is capable of simulating millions of sequences with length comparable with human chromosomes. Methods such as the Positional Burrows-Wheeler Transform (PBWT) [13], its ARG-based extension tree consistent PBWT [14], and tsinfer [15] can be used to efficiently process and store genomic sequences, but all of these approaches are designed for actively recombining organisms. Moreover, the primary population models underlying these methods are the Kingman coalescent [16], the Wright-Fisher model [17, 18] and the Li-Stephens model [19]. We recently developed approaches for compressing and accessing viral genealogies that dramatically reduce space and memory requirements [20,21], but there are no viral genealogy simulation methods that can efficiently produce pandemic-scale datasets.

Coalescent models are powerful tools for studying humans, many other eukaryotes, and pathogen populations (*e.g.* [22]). However, their assumptions are often violated in epidemiological settings. First, the effective population size is usually modelled either as piece-wise constant or as exponential growth. However, the coalescent with exponential growth and birth-death do not result in equivalent genealogies [23]. Second, it's not simple to use coalescent models to describe the effects of selection. If we consider the pandemic on a longer time period, basic birth-death models (e.g. [24]) are not an appropriate choice, since the reproductive rate usually decreases with time as collective immunity builds up or as the susceptible population is exhausted. These limitations are often addressed in epidemiology using compartmental models, such as SI, SIS and SIR [25], or their stochastic realisations, which are also birth-death processes.

Simulating realistic selection in backward-time models is a well-known challenging problem. A common workaround is to assume a single deterministic

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

frequency trajectory or to generate a stochastic frequency trajectory in forward time, and then to simulate the ancestry of the samples around the selected site in a coalescent framework (e.g., [26, 27]). However, more complex models of selection, including *e.g.,* gene-gene interactions, or epistasis, are often beyond the scope of such coalescent models. Nonetheless, epistasis is thought to be an important component of viral evolutionary processes [28, 29], and incorporating the effects of such complex evolutionary dynamics is essential for accurate simulations of evolution.

We introduce a novel simulation method that can rapidly generate pandemic-scale viral genealogies. Our approach is a forward-backward algorithm where we generate a series of stochastic events forward in time, then traverse backwards through this event series to generate the realized viral genealogy for a sample taken from the full population. Our framework includes the accumulation of immunity within host populations and of viral mutations that affect the fitness of descendant lineages. Our method is extremely fast, and can produce a phylogeny with 50 million total samples in just 88.5 seconds. The genealogies output from our simulation are compatible with phastSim [30], making it possible to generate realistic genome data for the simulated samples. This framework empowers efficient and realistic simulation of pandemic-scale viral datasets.

2 Design and Implementation

Our epidemiological model is a compartmental model [31] (see SI 1 for a brief introduction to compartmental models), and the realisations of the stochastic processes are drawn using the Gillespie algorithm [32]. The different compartments in our model are defined based on several interacting real-world complexities: (1) host population structure with corresponding population-specific viral frequencies and contact rates, (2) separate host infectious groups resulting from different viral haplotypes, and (3) different host susceptibility groups.

We break the simulation into two phases. In the first one (the forward pass), we generate a population-level epidemiological process which is represented as the series of events (Figure 1) resulting from the "reactions" (Table 1). These events then influence the properties of the viral genealogy, which is sampled in the second phase (the backward pass). The specific viral genealogy is sampled conditioned of the population-level epidemiological process using a coalescent framework.

Table 2 lists all the features which determine the simulation. In the beginning, the user should specify the number U of mutable sites (see Section 2.1), the number T of susceptibility types (Section 2.2), and the number K of populations (Section 2.3).

2.1 Mutations

Because this simulation framework focuses on generating the viral genealogy, and not genomes, we track only mutations at genome sites that have a large

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

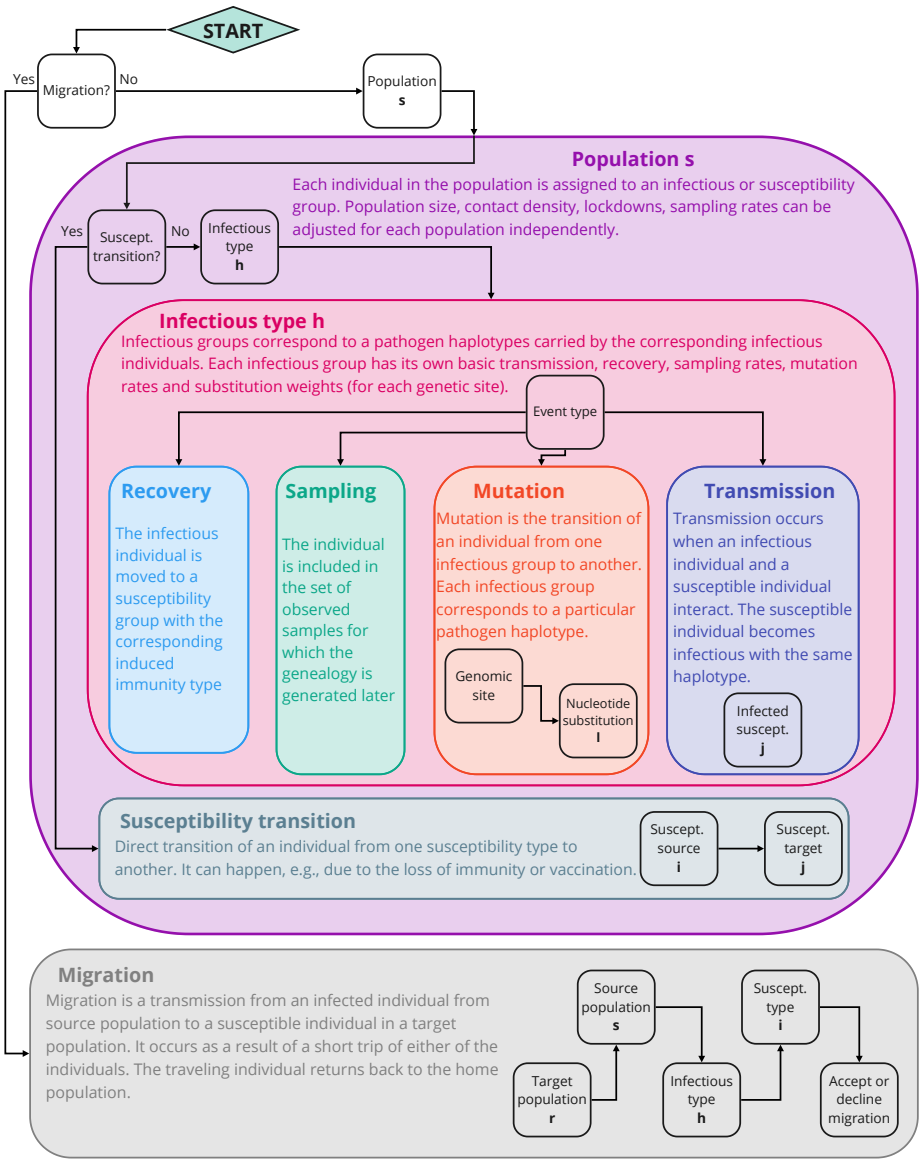

Figure 1: The scheme of the nested family Gillespie algorithm used to generate an event in the forward run. The corresponding reactions are listed in Table 1. Black squares correspond to the consecutive steps, where the subfamilies are chosen with the weights given by their propensities. The propensities for each step are cached and updated only if they change due to an event. For migration propensities, the rejection approach is used instead (SI 3).

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Reaction	Description	Number of reactions
$S_i^s, I_h^s \to I_h^s, I_h^s$	*Transmission* of haplotype h to an individual of susceptibility type i in population s	$4^U TK$
$I_h^s \to S_{i(h)}^s$	*Recovery* or *sampling* of an individual infected with haplotype h in population s. Susceptibility type $i(h)$ is determined by the haplotype.	$2 \cdot 4^U K$
$I_h^s \to I_l^s,\ h \neq l$	*Mutation* of haplotype h into haplotype l in population s as a result of a single nucleotide substitution.	$4^U 3U$
$S_i^s \to S_j^s,\ i \neq j$	*Susceptibility transition* of an individual with susceptibility type i to susceptibility type j (*e.g.* vaccination or loss of immunity) in population s.	$T(T-1)$
$S_i^r, I_h^s \to I_h^r, I_h^s,\ r \neq s$	*Migration* is a transmission of haplotype h from population s to an individual of susceptibility type i in population r.	$4^U TK(K-1)$

Table 1: The list of reactions and corresponding epidemiological events simulated by the Gillespie algorithm in our model, and the number of reactions in each category in function of the number of mutable sites U, number of susceptible individuals T, and the number of populations K.

positive effect on viral fitness. That is, these mutations enhance the transmissibility of the virus or lead to immunity escape. We expect this will typically be a relatively small number of mutations relative to the size of the viral genome, simplifying the problem. To efficiently model neutral genetic variation we suggest using `phastSim` [30] on a tree generated by our algorithm; the output produced by our method can be directly imported into phastSim for downstream processing.

To define the intended model of selection on new mutations, the user specifies the number U of mutable sites and their specific fitness effects (i.e., their effect on the birth rate). The mutations are modelled as single nucleotide substitutions, so each site has four possible variants (A, T, C and G). Mutations lead to the appearance of different haplotypes with different transmission and immunological properties. Up to $H = 4^U$ different haplotypes can appear in the simulation. Each haplotype h can be assigned its own specific U mutation rates $m^u(h)$ and $3U$ substitution probabilities $p_{h_1}^u$, $p_{h_2}^u$, $p_{h_3}^u$, one for each site u and for each of the 3 possible new nucleotides at site u. Transmission, recovery, and sampling rates, as well as mutation rates, susceptibility, and triggered sus-

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Model	Feature	Description	Value
Epidemiological model: every parameter can be set individually for each haplotype.	Transmission rate	The expected number of new infections per time unit caused by an individual infected with haplotype h if all the population were completely susceptible.	$\lambda_h \in (0;\infty)$
	Recovery rate	Rate at which an infectious individual becomes recovered after being infected with haplotype h.	$\rho_h \in (0;\infty)$
	Sampling rate	Rate at which an infectious individual is sampled after being infected with haplotype h.	$\zeta_h \in [0;\infty)$
	Mutation rate	Rate at which a genetic site u mutates. Can be set independently for each mutable site in function of haplotype h.	$m^u(h) \in (0;\infty)$
	Substitution probabilities	The probabilities of particular nucleotide substitution at haplotype h given that the mutation occurred at the site u.	$[p^u_{h_1}, p^u_{h_2}, p^u_{h_3}]$ $0 \leq p^u_{h_i} \leq 1$ $p^u_{h_1} + p^u_{h_2} + p^u_{h_3} = 1$
	Susceptibility	The multiplier which allows to change the relative susceptibility to haplotype h of an individual with susceptibility type i.	$\sigma_{ih} \in [0;\infty)$
	Susceptibility transition rate	The rate at which susceptible individuals move from one susceptibility type to another without being infected. This allows to model the loss of immunity with time or vaccination efforts.	$[0;\infty]$
Population model	Population size	Total number of individuals in population s.	$\mathbb{N}^s \in (0;\infty)$
	Contact density	The multiplicative modifier of transmission rate corresponding to the relative number of contacts in population s. It is used to describe differences in social behaviour of the host population (e.g. population density, wearing masks, lockdown effect).	$\delta^s \in [0;\infty)$
	Lockdown	Conditions to impose and lift lockdown in population k (determined by the proportion of infectious individuals in the population) and the contact density during the lockdown.	
	Sampling effort	This modifier increases or decreases the sampling rate in population s.	$c^s \in [0;\infty]$
	Migration probability	The probability that an individual from population s is temporarily visiting population r.	$\mu_{sr} \in [0;1]$

Table 2: List of features which determine the simulation scenario. All the rates are normalized by the number of individuals in a particular group (i.e. the number of individuals infected with a particular haplotype or individuals of a certain susceptibility type). The rates are measured in terms of events per time unit.

ceptibility types can be defined individually for every haplotype. Of particular interest, gene-gene, or epistatic, interactions can be flexibly modelled using this approach.

We refer to sequences carrying particular sets of variants as "haplotypes", because two identical sequences can appear as a results of different mutation events, so they might not belong to the same clade in the final tree.

2.2 Epidemiological model

To model the host immunity process, we use a generalised SI-model. The compartments within each population represent different types of susceptible individuals or infectious individuals.

Different susceptible compartments in the same host population are used to model different types of immunity. These can correspond for example to host individuals that have recovered from previous exposure to different viral haplotypes. Susceptible compartments can also be used to represent different vaccination statuses. For each susceptible compartment S_i, and for each viral haplotype h we consider a susceptibility coefficient σ_{ih} which multiplicatively changes the transmission (birth) rate of the corresponding haplotype. In particular, $\sigma_{ih} = 0$ corresponds to absolute resistance, similar to the R-compartment

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

in SIR-model, but specific to individuals who have immunity type i and are exposed to haplotype h. $0 < \sigma_{ih} < 1$ would correspond to partial immunity, while $\sigma_{ih} > 1$ corresponds to increased susceptibility.

Different infectious compartments within a host population correspond to individuals infected by a haplotype and can potentially infect susceptible hosts. As we mentioned in the section 2.1, the transmission rate λ_h, recovery rate ρ_h and mutation rates can be set independently for each haplotype h. After recovery, a host individual that was infected with haplotype h, and therefore was in compartment I_h^s for some population s, is moved to the corresponding susceptibility (immunity) compartment $S_{i(h)}^s$. Different haplotypes might however lead to the same types of immunity.

NB: The evolution of individual immunity is modeled as Markovian - it is determined only by the latest infection, and does not have memory about previous infections. Whether this assumption provides an accurate approximation of the immunity dynamics within the host population is an important consideration and may depend on the specific pathogen biology. Different haplotypes can lead to the same immunity. Some immunity types can be specific, *e.g.*, to vaccination immunity without being associated with any haplotypes at all.

The rate of transmission of viral lineages within a population also depends on how frequently two host individuals come in contact with each other. To flexibly accommodate such differences, each population s is assigned a contact density δ^s parameter. This parameter can be used to simulate differences in the local population density, social behaviours etc. The rate for an individual from susceptibility class S_i^s (the susceptible compartment i within population s) to be infected with haplotype h by another individual within population s is

$$|S_i^s|\sigma_{ih}\lambda_h|I_h^s|\delta^s/N^s,$$

where $|S_i^s|$ is the number of individuals in S_i^s, I_h^s is the class of individuals infected with haplotype h in population s, λ_h is the baseline transmission rate of individuals infected with haplotype h, and N^s is the total population size of deme s.

Direct transitions between susceptible compartments are possible, for which users can specify a transition matrix for susceptible compartments. A transition between susceptible compartment can be used for example to model a vaccination event, or the loss of immunity with time.

2.3 Population model

2.3.1 Demes

The population model is based on an island (demic) model. Each population is described at each point in time by its total size N^s, number $|I_h^s|$ of infectious individuals with each viral haplotype h, number $|S_i^s|$ of susceptible host individuals for each susceptibility type i, relative contact density δ^s, and a population-specific lockdown strategy and effectiveness.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2.3.2 Lockdown

Several governments have imposed lockdowns during the COVID-19 pandemic as an effort to control the spread of SARS-CoV-2. Understanding the effects of lockdowns is a crucial concern for designing effective public health strategies. We implement lockdowns as follows. When the total number of simultaneously infectious individuals in population s surpasses a certain user-defined population-specific percentage (e.g. 1%) of the population size N^s, the lockdown is imposed and the contact density δ^s is changed to a new (typically lower) lockdown- and population-specific value. When the percentage of the infectious individuals drops below a user-specified value (e.g. 0.1%) the lockdown is lifted and the contact density in population s reverts to its initial value δ^s.

2.3.3 Migration

Migration is described by a matrix μ_{sr} which defines the probabilities at which an individual from source population s can be found in target population r. In our model, new infections occur from the contact between infectious individuals from one population and susceptible individuals from the second population. It can be due to the travel of a susceptible individual to a source population, where it contracts an infection, and then returns back to the home population (first term in equation 1); or, to the travel of an infectious individual into a target population where this individual transmits the infection to a susceptible individual (second term in equation 1). The derivation of each term is similar to the derivation of within-population transmission (see SI equation 1). This model corresponds to short-term travel such as tourist or business trips, where an individual returns soon back to the home population. The proposed process is different from the traditional migration modelling in the population genetics, when an individual moves to a new population and remains there. The rate at which new infections of individuals with immunity i in population r are caused by haplotype h in population s is

$$M(r,i;s,h) = \lambda_h \sigma_{ih} \mu_{ss} \left(\mu_{rs} \delta^s |S_i^r| \frac{|I_h^k|}{N^s} + \mu_{sr} \delta^r \frac{|S_i^r|}{N^r} |I_h^s| \right), \tag{1}$$

where $\mu_{ss} = 1 - \sum_{q \neq s} \mu_{sq}$ is the probability that an individual originally from population s is currently not in a different population. Since it is computationally demanding to keep track of how each migration rate between each pairs of compartment is affected by each simulation event, instead we keep track of cumulative upper bounds on such migration rates (see SI 3 for details). In the case a potential migration event is sampled according to these upper bounds, we then proceed to calculate the precise migration rates and only sample a specific migration event according to its own exact rate. This is efficient when cross-population transmissions (migrations) are rare compared to within-population transmissions. This algorithmic implementation might perform suboptimally if population structure is extremely weak.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2.4 Sampling

Sampling is modelled using a continuous sampling scheme. In this scheme every individual infected with haplotype h in population s is sampled at rate $c^s\zeta_h$, the product of the haplotype-specific sampling rate ζ_h and the population modifier c^s. Sampled individuals then instantly recover and are moved to susceptible group $S^s_{i(h)}$, effectively increasing the recovery rate ρ_h for I^s_h by $c^s\zeta_h$. Alternatively, one can think about this sampling scheme as setting the recovery rate for I^s_h to $\rho_h + c^s\zeta_h$ and sampling an individual in I^s_h upon its recovery with probability $c^s\zeta_h/(\rho_h + c^s\zeta_h)$. More details can be found in supplementary information 7.

2.5 Algorithm

The simulation process is split into two parts, forward and backward. In the forward run, a chain of events (including sampled cases) describing the dynamics of the epidemiological process at the population level is generated with the Gillespie algorithm [32]. In the backward run, our method simulates the genealogy of the samples in a coalescent-like manner while conditioning on the events generated during the forward run.

2.5.1 Forward run

The forward run generates a chain of events which reflects the dynamics of the pandemic. Our implementation of the Gillespie algorithm is based on three algorithmic ideas: logarithmic direct method [32] (the events, or "reactions", are organised in nested families, Figure 1), rejection-based approach [33] for migrations (see SI 3 for details), and organising propensity dependencies to avoid updating those propensities which are not affected by events [34]. Details are given in SI 4.

2.5.2 Backward run

The backward run randomly builds a genealogical tree of the samples while conditioning on this chain of events generated in the forward run.

All of the ancestral lineages of the samples generated in the forward run belong to one of the infectious compartment corresponding to a haplotype h in a specific population s. Lineages are exchangeable within each compartment. Conditional on any event generated in the forward run, it is straightforward to calculate the probability that the event affected zero, one or two sample ancestral lineages in the backward run (see SI 5 for details).

2.5.3 Implementation details

`VGsim` provides a convenient Python user interface. Performance-critical parts are implemented in C++ via Cython [35]. The dependencies are kept to a

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

minimum: NumPy [36] and `mc_lib` —a small wrapper of the NumPy C API for generation of pseudorandom numbers in Cython [37].

3 Results

3.1 Forward run performance

To test the scalability of the population model, we performed simulations with $K = 2, 5, 10, 20, 50$ and 100 total host populations with $2 \cdot 10^9/K$ individuals in each and generated 10 million events (see Figure 1) in each run (see Table 3). There are 16 haplotypes resulting from two segregating sites with mutation rates 0.01 in each of them (this is unrealistically high, but it ensures that all the haplotypes appear in the simulation), and three susceptibility group with the first group corresponding to the absence of immunity, the second group corresponding to partial immunity and the last one corresponding to resistance to all strains. The transmission rate is $\lambda = 2.5$ for all haplotypes except one, and $\lambda = 4.0$ for this last haplotype. The recovery rate is $\rho = 0.9$, the sampling rate is $\zeta = 0.1$ (so, the effective reproductive number is 2.5 which approximately correspond to SARS-CoV-2 [38] if the time unit is interpreted as one week). All the migration probabilities were set to $\mu = M/(K-1)$, where M is the cumulative migration rate from a population. The runtime of the forward algorithm does not depend only on the cumulative migration rate M, but also on the percentage of potential migrations rejected by the algorithm (see Section 2.3.3 for details), which appears to grow with M. However, the effect on runtime is relatively modest (in contrast to the naive algorithm which is quadratic in the number of populations, see Table 1) indicating that this approach scales well to pandemic-scale simulations.

3.2 Backward run and overall performance

Our implementation of the backward run algorithm relies on an efficient and compact tree representation, a Prüfer-like code [39]. Each node is associated with an index in an array, and the corresponding entry in the array is the index of the parent node. The time needed to generate a tree mainly depends on two factors: the total number of events generated in the forward run, and the total number of samples in a tree. We report the execution time of the backward run in Table 4. The combined approach is sufficiently fast that it can be used to generate many replicate simulations as is often required to validate empirical methods and to train model parameters. Table 4 also shows the forward time, the total number of generated events and the total number of infected individuals over the simulation for various sampling rate (where the sampling rate $\zeta = 0.01$ is 1 in 100 cases is sampled, $\zeta = 0.1$ corresponds to 1 in 10 cases is sampled, and $\zeta = 1.0$ means that every case is sampled), and various sample sizes. The simulation assumes the absence of immunity after infection (SIS-model), which allows to run the simulation sufficiently long to collect enough samples (instead,

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Cumulative migration probability M	Number of demes K					
	2	5	10	20	50	100
0.001	28.7s 0.09%	30.0s 0.12%	31.9s 0.11%	35.1s 0.11%	47.2s 0.11%	69.3s 0.12%
0.002	29.2s 0.17%	30.4s 0.21%	32.3s 0.16%	35.3s 0.2%	47.0s 0.21%	70.1s 0.2%
0.005	29.4s 0.33%	30.7s 0.51%	32.5s 0.25%	35.6s 0.46%	48.1s 0.52%	70.4s 0.43%
0.01	29.4s 0.75%	30.6s 1.16%	32.9s 0.85%	35.5s 0.78%	48.0s 0.93%	70.9s 0.95%
0.1	30.3s 2.04%	31.9s 6.56%	33.8s 6.15%	37.0s 6.94%	50.3s 5.98%	73.0s 5.08%

Table 3: Run time to generate 10 million events. The second number is the percentage of discarded events (due to migration acceptance/rejection). There are $16 = 4^2$ haplotypes and 3 susceptible compartments. The sampling rate is set to $\zeta = 0.1$, recovery rate is $\rho = 0.9$, transmission rate is $\lambda = 2.5$. The tests were run on a server node with Intel Xeon Gold 6152 2.1-3.7 GHz processor and 1536GB of memory.

with an SIR-model susceptible individuals can be exhausted before the desired number of samples is simulated).

To showcase the limit of applicability of our simulator, we also show in Table 4 the computational demand for the simulation of an unrealistically large (for now) genealogy of 150 million samples (with 1 in 100 cases sampled), for which we almost reached the memory limit available on our supercomputer node (1536GB) [40]. The total number of infections in the population is more than 15 billion cases, with the total number of events being more than 30 billions. The forward run time was approximately 9.5 hours and the backward run time was 13.5 minutes.

3.3 Comparison with other simulators

There are many epidemiological simulators which are capable of producing viral genealogies. Agent-based simulators (e.g. `nosoi` [41], `FAVITES` [42]) allow to create very detailed models, because every agent's parameters can be set individually. The trade-off is they are computationally demanding, so only relatively small scenarios can be modelled. Other simulators (e.g. `MASTER` [43] and `TiPS` [44]) implement Gillespie algorithm for compartmental models, but they currently lack a simple user interface, instead requiring users to specify the full set of reaction equations, and they might be not specifically optimised for epidemiological purposes. On the other hand, both `MASTER` and `TiPS` implement

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Sampling rate		Sample size (number of tree leaves)					
		10^5	10^6	$5 \cdot 10^6$	10^7	$5 \cdot 10^7$	$1.5 \cdot 10^8$
0.01	Forward time	27.84s	290.86s (4min 50.86s)	1275.53s (21min 15.53s)	2487.73s (41min 27.73s)	11295.01s (3h 8m 15.01s)	34558.86s (9h 35m 58.86s)
	Backward time	0.85s	7.44s	26.93s	50.27s	217.51s (3min 37.51s)	813.25s (13min 33.25s)
	Memory	1.67MB	10.87GB	49.54GB	94.64GB	442.69GB	1.34TB
	Total number of generated events	34,038,092	286,381,088	1,120,365,070	2,121,897,004	9,878,131,708	30,152,423,891
	Total number of infections	24,040,769	185,954,943	619,559,504	1,119,957,985	4,994,200,627	15,121,211,248
0.1	Forward time	2.18s	29.89s	154.15s (2min 34.15s)	296.43s (4min 56.43s)	1283.01s (21min 23.01s)	3470.47s (57min 50.47s)
	Backward time	0.1s	0.96s	4.68s	8.99s	34.2s	90.29s (1min 30.29s)
	Memory	1.68MB	1.68MB	5.51GB	12.5GB	53.27GB	143.32GB
	Total number of generated events	3,491,562	34,125,248	155,922,768	285,874,161	1,120,657,092	3,122,658,422
	Total number of infections	2,489,943	24,101,573	105,814,516	185,656,462	619,705,716	1,619,831,406
1.0	Forward time	0.23s	2.2s	13.63s	30.32s	154.99s (2min 34.99s)	405.39s (6min 45.39s)
	Backward time	0.01s	0.15s	0.92s	2.08s	11.35s	32.48s
	Memory	1.67MB	1.68MB	1.66MB	1.67MB	5.54GB	20.9GB
	Total number of generated events	350,517	3,492,789	17,271,140	34,113,125	155,899,482	401,912,500
	Total number of infections	250,290	2,490,805	12,261,217	24,093,104	105,799,613	251,613,148

Table 4: Run time in seconds to generate a random genealogy for a sample of a certain size for different sampling rates. The execution time is shown split into the time demand for the forward run and the one for the backward run only. We simulated $16 = 4^2$ haplotypes and no host immunity. The recovery rate is $\rho = 1.0 - \zeta$, with ζ the sampling rate, while the transmission rate is $\lambda = 2.5$ for all 16 haplotypes. The tests were run on a server node with an Intel Xeon Gold 6152 2.1-3.7 GHz processor and 1536GB of memory.

approximate methods (tau-leaping and hybrid approaches), which decrease the time of forward simulation by orders of magnitude and hence might outperform our simulator depending on simulation scenario. `VGsim` is optimised to scale for large epidemics and genealogies, though approximate approaches are not available in the current implementation. It also has a simple and flexible user interface which helps merge together several complexities (epidemiology, evolution, population structure and cross-immunity). The detailed discussion of different simulating frameworks and detailed comparisons with them can be found in SI 8.

3.4 Simulating realistic nucleotide mutations

Our simulation framework generates a phylogenetic tree, and if the user specifies a scenario with strongly selected mutations, these are included in the output; we, however, do not include a method for simulating many neutral variants. To facilitate studies that require full viral genome sequences we have made the output of our approach compatible with that of `phastSim` [30]. Briefly, a user can easily load the output of our software into `phastSim`, and `phastSim` will generate neutral mutations, while leaving previously determined selected

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

mutations unaffected.

4 Availability and Future Directions

`VGsim` is freely available from `https://github.com/Genomics-HSE/VGsim` under GPL-3.0 License. It is tested for Python 3.6 and later under Ubuntu and macOS. The documentation and tutorials are published at `https://vg-sim.readthedocs.io/`.

The future development of `VGsim` will include the following updates. We will consider improving performance by adding the τ-leaping algorithm and optimizing memory usage to handle larger numbers of genetic sites. We will also extend the available models by adding super-spreading events, life-cycle compartments, and new sampling schemes. We will also add recombination events, though they seem to be relatively rare [45] and so far are not a major driver of SARS-CoV-2 genetic diversity and evolution.

`VGsim` is particularly useful for simulating large datasets, in particular, in those cases when agent-based simulators become inefficient (see SI for more detailed discussion 8). It is primarily optimised for the studies of world-wide pandemic scenarios, and it is motivated by the features of the ongoing SARS-CoV-2 pandemic. We plan for the future to add more features which would generalise its applicability to different pathogens (e.g. with complex life-cycle). Further possible optimisations of our algorithm will also be investigated.

Our implementation allows simulations of scenarios with a few loci with strong phenotypic effects. However, we cannot simulate the effect of many loci with mild fitness effects. While mild and widespread fitness effects can be simulated by phastSim, they are simulated in a typical phylogenetic way (using a substitution codon matrix with specifiable nonsynonymous/synonymous ratios) and so their impact on the tree shape and epidemiological dynamics are neglected.

5 Conclusion

We developed a fast simulator `VGsim` which can be used to produce genealogies of millions of samples from world-scale pandemic scenarios. Our method models many major aspects of epidemiological dynamics: viral molecular evolution, host population structure, host immunity, vaccinations and lockdowns. We expect that `VGsim` will be a useful tool in method validation and in simulation-based statistical inference.

The performance of our simulator should meet the performance requirements of most studies. The flexible Python API, combination of epidemiological (including cross-immunity), population and evolutionary models make it a timely tool for the modern and future research.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

6 Acknowledgment

We are thankful to Mikhail Shishkin for testing VGsim on the Apple Silicon M1 processor. VSh, VSp, DS, RCD were funded within the framework of the HSE University Basic Research Program. EB acknowledges support within the Project Teams framework of MIEM HSE. VSh was supported by RFBR grant 20-04-60556 while working on section 2.3. This research was supported in part through computational resources of HPC facilities at NRU HSE. RCD was supported in part by NIH/NIGMS R35GM128932. NDM was supported by the European Molecular Biology Laboratory.

7 Data availability statement

There is no data and reagent used in the paper. The code is available at the GitHub repository associated with this project:
`https://github.com/Genomics-HSE/VGsim`.

References

[1] Hodcroft EB, De Maio N, Lanfear R, MacCannell DR, Minh BQ, Schmidt HA, et al.. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. Nature Publishing Group; 2021.

[2] Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. Science. 2020;369(6501):297-301. Available from: `https://science.sciencemag.org/content/369/6501/297`.

[3] Nadeau SA, Vaughan TG, Scire J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. Proceedings of the National Academy of Sciences. 2021;118(9). Available from: `https://www.pnas.org/content/118/9/e2012008118`.

[4] Ladner JT, Larsen BB, Bowers JR, Hepp CM, Bolyen E, Folkerts M, et al. An Early Pandemic Analysis of SARS-CoV-2 Population Structure and Dynamics in Arizona. mBio. 2020;11(5). Available from: `https://mbio.asm.org/content/11/5/e02107-20`.

[5] Komissarov AB, Safina KR, Garushyants SK, Fadeev AV, Sergeeva MV, Ivanova AA, et al. Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia. Nature Communications. 2021 Jan;12(1):649. Available from: `https://doi.org/10.1038/s41467-020-20880-z`.

[6] Lycett SJ, Hughes J, McHugh MP, da Silva Filipe A, Dewar R, Lu L, et al. Epidemic waves of COVID-19 in Scotland: a genomic perspective on

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the impact of the introduction and relaxation of lockdown on SARS-CoV-2. medRxiv. 2021. Available from: https://www.medrxiv.org/content/early/2021/01/20/2021.01.08.20248677.

[7] Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. Nature Medicine. 2021 Mar;27(3):440-6. Available from: https://doi.org/10.1038/s41591-021-01255-3.

[8] Garcia-Beltran WF, Lam EC, St Denis K, Nitido AD, Garcia ZH, Hauser BM, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell. 2021. Available from: https://www.sciencedirect.com/science/article/pii/S0092867421002981.

[9] Burioni R, Topol EJ. Assessing the human immune response to SARS-CoV-2 variants. Nature Medicine. 2021 Apr;27(4):571-2. Available from: https://doi.org/10.1038/s41591-021-01290-0.

[10] Zeng HL, Dichio V, Rodríguez Horta E, Thorell K, Aurell E. Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes. Proceedings of the National Academy of Sciences. 2020;117(49):31519-26. Available from: https://www.pnas.org/content/117/49/31519.

[11] Rochman ND, Wolf YI, Faure G, Mutz P, Zhang F, Koonin EV. Ongoing global and regional adaptive evolution of SARS-CoV-2. Proceedings of the National Academy of Sciences. 2021;118(29). Available from: https://www.pnas.org/content/118/29/e2104241118.

[12] Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology. 2016 05;12(5):1-22. Available from: https://doi.org/10.1371/journal.pcbi.1004842.

[13] Durbin R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). Bioinformatics. 2014 01;30(9):1266-72. Available from: https://doi.org/10.1093/bioinformatics/btu014.

[14] Shchur V, Ziganurova L, Durbin R. Fast and scalable genome-wide inference of local tree topologies from large number of haplotypes based on tree consistent PBWT data structure. bioRxiv. 2019. Available from: https://www.biorxiv.org/content/early/2019/02/06/542035.

[15] Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nature Genetics. 2019 Sep;51(9):1330-8. Available from: https://doi.org/10.1038/s41588-019-0483-y.

[16] Kingman JFC. On the genealogy of large populations. Journal of Applied Probability. 1982;19(A):27–43.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[17] Fisher RA, Russell EJ. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character. 1922;222(594-604):309-68. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1922.0009.

[18] Wright S. EVOLUTION IN MENDELIAN POPULATIONS. Genetics. 1931;16(2):97-159. Available from: https://www.genetics.org/content/16/2/97.

[19] Li N, Stephens M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. Genetics. 2003;165(4):2213-33. Available from: https://www.genetics.org/content/165/4/2213.

[20] Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nature Genetics. 2021 Jun;53(6):809-16. Available from: https://doi.org/10.1038/s41588-021-00862-7.

[21] McBroome J, Thornlow B, Hinrichs AS, De Maio N, Goldman N, Haussler D, et al. matUtils: Tools to Interpret and Manipulate Mutation Annotated Trees. bioRxiv. 2021. Available from: https://www.biorxiv.org/content/early/2021/04/04/2021.04.03.438321.

[22] De Maio N, Wilson DJ. The Bacterial Sequential Markov Coalescent. Genetics. 2017 05;206(1):333-43. Available from: https://doi.org/10.1534/genetics.116.198796.

[23] Lambert A, Stadler T. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. Theoretical Population Biology. 2013;90:113-28. Available from: https://www.sciencedirect.com/science/article/pii/S0040580913001056.

[24] Stadler T. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. Journal of Theoretical Biology. 2009;261(1):58-66. Available from: https://www.sciencedirect.com/science/article/pii/S0022519309003300.

[25] Brauer F. Compartmental models in epidemiology. In: Mathematical epidemiology. Springer; 2008. p. 19-79.

[26] Kern AD, Schrider DR. Discoal: flexible coalescent simulations with selection. Bioinformatics. 2016 08;32(24):3839-41. Available from: https://doi.org/10.1093/bioinformatics/btw556.

[27] Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

locus. Bioinformatics. 2010 06;26(16):2064-5. Available from: https://doi.org/10.1093/bioinformatics/btq322.

[28] Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. PLOS Genetics. 2011 02;7(2):1-11. Available from: https://doi.org/10.1371/journal.pgen.1001301.

[29] Sanjuán R, Moya A, Elena SF. The contribution of epistasis to the architecture of fitness in an RNA virus. Proceedings of the National Academy of Sciences. 2004;101(43):15376-9. Available from: https://www.pnas.org/content/101/43/15376.

[30] De Maio N, Weilguny L, Walker CR, Turakhia Y, Corbett-Detig R, Goldman N. phastSim: efficient simulation of sequence evolution for pandemic-scale datasets. bioRxiv. 2021. Available from: https://www.biorxiv.org/content/early/2021/03/16/2021.03.15.435416.

[31] Kermack William Ogilvy MAG, Thomas WG. Thomas A contribution to the mathematical theory of epidemics. Proceedings of Royal Society A. 1927;115:700 721.

[32] Gillespie DT. Stochastic Simulation of Chemical Kinetics. Annual Review of Physical Chemistry. 2007;58(1):35-55. PMID: 17037977. Available from: https://doi.org/10.1146/annurev.physchem.58.032806.104637.

[33] Thanh VH, Priami C, Zunino R. Efficient rejection-based simulation of biochemical reactions with stochastic noise and delays. The Journal of Chemical Physics. 2014;141(13):134116. Available from: https://doi.org/10.1063/1.4896985.

[34] Cao Y, Li H, Petzold L. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. The Journal of Chemical Physics. 2004;121(9):4059-67. Available from: https://doi.org/10.1063/1.1778376.

[35] Behnel S, Bradshaw R, Citro C, Dalcin L, Seljebotn DS, Smith K. Cython: The best of both worlds. Computing in Science & Engineering. 2011;13(2):31-9.

[36] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020 Sep;585(7825):357-62. Available from: https://doi.org/10.1038/s41586-020-2649-2.

[37] Burovski E, Godyaev D, Gorbunova V. mc_lib: Assorted small utilities for MC simulations with Cython;. Available from: https://doi.org/10.5281/zenodo.5169027.

medRxiv preprint doi: https://doi.org/10.1101/2021.04.21.21255891; this version posted December 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[38] Billah MA, Miah MM, Khan MN. Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. PLOS ONE. 2020 11;15(11):1-17. Available from: https://doi.org/10.1371/journal.pone.0242128.

[39] Prüfer H. Neuer Beweis eines Satzes über Permutationen. Arch Math Phys. 1918.

[40] Kostenetskiy PS, Chulkevich RA, Kozyrev VI. HPC Resources of the Higher School of Economics. Journal of Physics: Conference Series. 2021 jan;1740:012050. Available from: https://doi.org/10.1088/1742-6596/1740/1/012050.

[41] Lequime S, Bastide P, Dellicour S, Lemey P, Baele G. nosoi: A stochastic agent-based transmission chain simulation framework in R. Methods in Ecology and Evolution. 2020;11(8):1002-7. Available from: https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13422.

[42] Moshiri N, Ragonnet-Cronin M, Wertheim JO, Mirarab S. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences. Bioinformatics. 2018 11;35(11):1852-61. Available from: https://doi.org/10.1093/bioinformatics/bty921.

[43] Vaughan TG, Drummond AJ. A Stochastic Simulator of Birth–Death Master Equations with Application to Phylodynamics. Molecular Biology and Evolution. 2013 03;30(6):1480-93. Available from: https://doi.org/10.1093/molbev/mst057.

[44] Danesh G, Saulnier E, Gascuel O, Choisy M, Alizon S. Simulating trajectories and phylogenies from population dynamics models with TiPS. bioRxiv. 2020. Available from: https://www.biorxiv.org/content/early/2020/11/09/2020.11.09.373795.

[45] Turkahia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, Ye C, et al. Pandemic-Scale Phylogenomics Reveals Elevated Recombination Rates in the SARS-CoV-2 Spike Region. bioRxiv. 2021. Available from: https://www.biorxiv.org/content/early/2021/08/05/2021.08.04.455157.